FRONTIERS IN GENETICS

ORIGINAL RESEARCH ARTICLE
published: 14 November 2014
doi: 10.3389/fgene.2014.00399

Quantifying the impact of inter-site heterogeneity on the distribution of ChIP-seq data

***Jonathan Cairns*[1,2]*, *Andy G. Lynch*[2] *and Simon Tavaré*[2]**

[1] *Nuclear Dynamics Group, The Babraham Institute, Cambridge, UK*
[2] *Cancer Research UK Cambridge Institute, University of Cambridge, Cambridge, UK*

Edited by:
Mark D. Robinson, University of Zurich, Switzerland

Reviewed by:
Helen Piontkivska, Kent State University, USA
Thiruvarangan Ramaraj, National Center for Genome Resources, USA
Yaomin Xu, Vanderbilt University, USA

***Correspondence:**
Jonathan Cairns, Nuclear Dynamics Group, The Babraham Institute, Babraham Research Campus, Cambridge, CB22 3AT, UK
e-mail: jonathan.cairns@babraham.ac.uk

Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) is a valuable tool for epigenetic studies. Analysis of the data arising from ChIP-seq experiments often requires implicit or explicit statistical modeling of the read counts. The simple Poisson model is attractive, but does not provide a good fit to observed ChIP-seq data. Researchers therefore often either extend to a more general model (e.g., the Negative Binomial), and/or exclude regions of the genome that do not conform to the model. Since many modeling strategies employed for ChIP-seq data reduce to fitting a mixture of Poisson distributions, we explore the problem of inferring the optimal mixing distribution. We apply the Constrained Newton Method (CNM), which suggests the Negative Binomial - Negative Binomial (NB-NB) mixture model as a candidate for modeling ChIP-seq data. We illustrate fitting the NB-NB model with an accelerated EM algorithm on four data sets from three species. Zero-inflated models have been suggested as an approach to improve model fit for ChIP-seq data. We show that the NB-NB mixture model requires no zero-inflation and suggest that in some cases the need for zero inflation is driven by the model's inability to cope with both artifactual large read counts and the frequently observed very low read counts. We see that the CNM-based approach is a useful diagnostic for the assessment of model fit and inference in ChIP-seq data and beyond. Use of the suggested NB-NB mixture model will be of value not only when calling peaks or otherwise modeling ChIP-seq data, but also when simulating data or constructing blacklists *de novo*.

Keywords: ChIP-seq, Negative Binomial, mixture model, zero-inflation, high-throughput sequencing

1. INTRODUCTION

Chromatin Immunoprecipitation followed by sequencing (ChIP-seq) is an experiment for the genome-wide location of events such as transcription factor binding sites and histone modifications (Park, 2009; Cairns et al., 2013). These events provide information on chromatin status, a topic of great interest in epigenetics. As with all sequencing assays, ChIP-seq can be represented as count data across the genome. Commonly, researchers need algorithms that find sites where the counts are larger than one would expect under a null noise model, a procedure known as "peak-calling."

Consider read counts x_i, where i indexes over genomic bins, for a single ChIP-seq sample. When modeling these counts as independent samples from some Poisson random variable X, a common problem is overdispersion—that is, the property that the observations have greater variance than allowed for by the model. In our case, the Poisson distribution requires that mean and variance are equal, an assumption that (even when ignoring between-sample variability) is violated by ChIP-seq data, impairing model fit (Spyrou et al., 2009).

The choice of distribution for X is important when peak-calling; in particular, underestimating the variance should decrease specificity. Indeed, many have commented on the presence of false positives in peak-calling and how this quantity varies depending on the choice of peak-caller (Landt et al., 2012).

A number of strategies exist to account for the extra variance. For example, we can allow the Poisson distribution to have site-specific means—that is, $X_i \sim Pois(\lambda_i)$. This strategy requires some sort of smoothing criterion to make parameter estimation robust (Zhang et al., 2008a). It is also difficult to expand this model to account for between-sample heterogeneity.

Many researchers use more general distributions—for example, the Negative Binomial (NB) model is used by Spyrou et al. (2009); Wu et al. (2010); Cairns et al. (2011); Song and Smith (2011) and others. The log-normal Poisson model has been used to model data from other high-throughput sequencing experiments, such as Serial Analysis of Gene Expression (SAGE) (Thygesen and Zwinderman, 2006).

Other strategies involve regression. ZINBA (Rashid et al., 2011) uses an NB model whose mean is regressed against known covariates, though the dispersion is fixed. Such a strategy requires knowledge of all appropriate dependent variables to capture the full variability.

An increasingly common approach is to use blacklists (Myers et al., 2011)—regions that have unusual mappability and previously have been seen to accumulate artifacts across many

ChIP-seq experiments. Reads that fall in blacklisted regions are removed.

Use of blacklists appears to improve peak-calling (Carroll et al., 2014), but has a number of downsides. Firstly, this strategy requires an organism-specific blacklist, and it is possible that the blacklist is non-exhaustive. Secondly, there is no reason to assume that enrichment cannot occur in these loci—a better alternative would be to model these regions appropriately, or perhaps separately. Thirdly, there may be sample-specific or copy-number specific events. For example, a transfected vector may artificially inflate copy number at a particular locus.

We consider the general form of the above models, taking a completely unsupervised random-effects approach. It is reasonable to retain the Poisson element in our model, since it has been shown that counts from a single site in a single library, sequenced repeatedly, follow a Poisson distribution (Marioni et al., 2008). However, by expanding to a mixture model setting, where x_i is a sample from $X_i \sim Pois(\Lambda)$ and the latent mixing variable Λ satisfies $\Lambda \sim f(\lambda)$, we can use $f(\lambda)$ to capture the genome-wide variation in our sample. Indeed, if Λ has gamma distribution, then X_i has NB distribution.

However, there is no biological or statistical reason, other than convenience, to believe that Λ is best modeled with the gamma distribution. We assess the fit of the NB model to the data; that is, investigate the appropriateness of the gamma distribution as a mixing distribution. Next, we use CNM to make inferences about the distribution of Λ, and see that it suggests a mixture of two distributions. Thus, we consider the NB-NB distribution and show that it provides a better fit to the data. We also consider zero inflation, and show that the poor fit of the NB distribution can be mistaken for the presence of zero-inflation in the data.

1.1. DATA

We use four different data sets, all of which are input samples (that is, untreated data). We focus on input samples because, in order to detect sites where counts differ from noise, it is important to model the noise appropriately. However, the methods used can also be applied to treated ChIP-seq data—see Supplementary Material. Samples were chosen to represent a variety of species. In each case, we considered only the first chromosome to avoid inter-chromosome normalization issues (Schmidt et al., 2010; Ross-Innes et al., 2012).

Sample	Species	Description	Read lengths	# Reads	Aligner
A	*C. Familiaris*	Normal liver tissue	45	480,843	MAQ
B	*M. Musculus*	Normal liver tissue	36	542,021	MAQ
C	*H. Sapiens*	Cancer cell line MCF7	44	633,931	BWA 0.5.5
D	*H. Sapiens*	Tumor sample BT82277	44	1,398,520	BWA 0.5.5

We partition the first chromosome into 100 bp bins, avoiding conservatively-chosen regions that span the telomeres and centromeres, as obtained from the UCSC Genome Browser at http://genome.ucsc.edu/. We did not remove duplicates—however, we found that deleting duplicates did not substantially affect our results (see Supplementary Material).

Code and data to reproduce the analysis are linked to in the Supplementary Material section.

2. INITIAL MIXTURE MODELS

2.1. POISSON

The Poisson model's Maximum Likelihood Estimate (MLE) occurs when its expected mean is equated to the observed mean: $\hat{\mu} = \bar{x}$.

2.2. NEGATIVE BINOMIAL

The MLE for the Negative Binomial (NB) distribution does not have closed form. However, we can fit the model using standard techniques, using the BFGS method implemented as *fitdistr()* in the *MASS* R package (Venables and Ripley, 2002).

Though the MLE for the dispersion of the NB distribution is biased, the bias is of the order $1/n$ (Saha and Paul, 2005) and we have enough bins in our data sets for the bias to be negligible.

2.3. LOG-NORMAL POISSON

Here, Λ is a log-normal random variable: $\log(\Lambda) \sim N(\mu, \sigma^2)$. The full log-likelihood involves a complicated integral:

$$\ell(\mu, \sigma^2) = \sum_{i=1}^{n} \log \int_{\mathbb{R}^+} \frac{1}{\lambda\sqrt{2\pi\sigma^2}} \frac{\exp\left(-\frac{(\log\lambda-\mu)^2}{2\sigma^2} - \lambda\right)\lambda^{x_i}}{x_i!} d\lambda$$

and the MLEs have no known closed form. Therefore, to estimate the parameters of this distribution, we take the following Bayesian approach:

1. Assume a weak prior distribution for each of μ (the mean) and σ^{-2} (the precision). We use conjugate prior distributions:

$$\mu \sim N(0, 10)$$
$$\sigma^{-2} \sim \Gamma(shape = 1, rate = 0.1)$$

   The prior distributions should have very little effect on the posterior, given that we have so many data.
2. Use the Metropolis-Hastings algorithm to sample from each parameter's posterior distribution.
3. Excluding a suitably-chosen burn-in period, take the mean of the posterior samples as an approximation to the maximum a posteriori (MAP) estimate.

We perform the analysis in WinBUGS (Lunn et al., 2000).

2.4. INITIAL RESULTS

Figure 1 shows an example of the fit of the above distributions to sample A. We see that the empirical distribution has a large tail that the fitted distributions cannot account for, which in turn affects their ability to model bins with low counts.

To explore this effect further, we investigate the properties of the underlying mixing distribution.

3. MIXTURE INFERENCE

As described in the introduction, suppose that we have multiple samples $\{x_i; i = 1, \ldots, n\}$ from a variable X distributed according to a Poisson mixture—that is,

$$X_i|\Lambda_i \sim Pois(\Lambda_i)$$
$$\Lambda_i \sim f(\lambda)$$

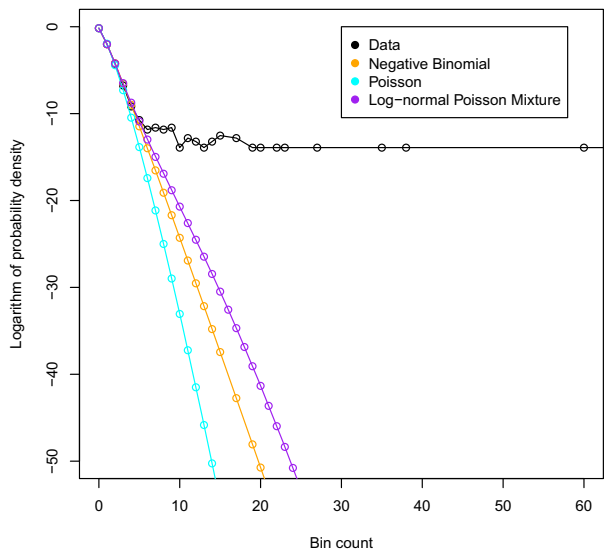

FIGURE 1 | Illustrating the best fits of the commonly used distributions to the count data for sample A. Note that none of these distributions can model adequately the heavy tail of large bin counts.

Our aim is to estimate the density $f(\lambda)$ from our observations $\{x_i\}$.

A formulation of the general problem, where $X_i|\Lambda_i$ has arbitrary distribution, is given in Roueff and Rydén (2005). In the case of the Poisson distribution, it is known that $\hat{f}(\lambda)$, the maximum likelihood estimator (MLE) of $f(\lambda)$, consists of a finite number of point masses: that is, if we adopt the MLE $\hat{f}(\lambda)$, then Λ is a discrete random variable, taking values from some vector $\boldsymbol{\theta}$ with associated probabilities $\boldsymbol{\pi}$, where $\boldsymbol{\theta}$ and $\boldsymbol{\pi}$ both have length L. It is also known that L, the number of support points used in the MLE distribution, can never exceed the number of distinct values in our X samples (Laird, 1978).

A number of methods have been proposed to infer the parameters L, $\boldsymbol{\theta}$, and $\boldsymbol{\pi}$, including those of Tucker (1963); Boes (1966); Simar (1976); Laird (1978). Here, we use the Constrained Newton Method (CNM), described in Wang (2007).

The CNM algorithm takes an initial value for $(L, \boldsymbol{\theta}, \boldsymbol{\pi})$, then repeats the following steps until convergence:

1. **Update** $(L, \boldsymbol{\theta})$, as follows: Suppose that $G(\lambda)$ is our current estimate for the mixing distribution, corresponding to the current value of $(L, \boldsymbol{\theta}, \boldsymbol{\pi})$. Now, consider some new candidate support point $\theta^\star$, and let $H_{\theta^\star}(\lambda)$ be the distribution that has a point mass at $\theta^\star$—in other words, $H_{\theta^\star}$ is the Dirac delta function $H_{\theta^\star}(\lambda) = \delta(\lambda - \theta^\star)$. Consider the "gradient function," the directional derivative

$$d(\theta^\star; G) = \frac{\partial \ell}{\partial \epsilon}\{(1-\epsilon)G + \epsilon H_{\theta^\star}\}$$

which we can think of as the rate at which the likelihood increases, as we shift probability mass across to the new support point $\theta^\star$.
The value of $\theta^\star$ that maximizes $d(\theta^\star; G)$ is added to $\boldsymbol{\theta}$, and L is increased by one. If multiple values of $\theta^\star$ maximize $d(\theta^\star; G)$, then we add each of them to $\boldsymbol{\theta}$, increasing L by one for each value.
2. **Update** $\boldsymbol{\pi}$ by calculating the MLE for $\boldsymbol{\pi}$ given $(\boldsymbol{\theta}, L)$. Wang (2007) show that this problem is equivalent to

$$\boldsymbol{\pi} = \min_{\boldsymbol{\pi}'} \|\boldsymbol{S}\boldsymbol{\pi}' - 2 \times \mathbf{1}\|^2$$

subject to $\sum_{i=1}^{L} \pi_i = 1$, $\boldsymbol{\pi}' \geq 0$, and where $S_{ij} = \frac{\partial}{\partial \pi_i}\ell_j$ (here, ℓ_j refers to the likelihood for a single data point, x_j). This minimization problem is solved using the Constrained Newton (CN) method.
3. If $\pi_i = 0$ for some i^*, then delete π_{i^*} and θ_{i^*} and reduce L by one.

Wang (2007) prove theoretical convergence and demonstrate that, in practice, convergence is dramatically faster than previous algorithms designed to find L, $\boldsymbol{\theta}$, and $\boldsymbol{\pi}$.

CNM reported that it did not converge for data sets C or D, and we check the output in the next section. In practice, we found that the value of L ranged from around 4 to 11 in our input samples.

3.1. DISTRIBUTION RECOVERY

First, we show that the CNM estimate $\hat{f}(\lambda)$ is appropriate for our data, by demonstrating that the empirical distribution of X can be recovered from the CNM estimate according to the following procedure:

1. Calculate the empirical probability density $\tilde{f}(x)$ directly, from the data.
2. Calculate the mixing distribution MLE $\hat{f}(\lambda)$ from the data according to the CNM algorithm.
3. Estimate the density $f(x)$ from the mixing distribution according to

$$\hat{f}(x) = \sum_{i=1}^{L} \pi_i f_P(x; \theta_i)$$

where $f_P(x; \theta_i)$ is the density of the Poisson distribution with mean θ_i.

If the CNM algorithm retains the key features of the true distribution $f(x)$, then we expect $\hat{f}(x)$ to be "similar" to the empirical distribution $\tilde{f}(x)$. The results of this comparison are shown in **Figure 2**.

3.2. SMOOTHING

The MLE $\hat{f}(\lambda)$ is desirable in the sense that it is the "best fit" to our data. However, it does not make physical sense to use $\hat{f}(\lambda)$ when modeling, which would make Λ a discrete variable. In reality, drawing from Λ represents a complicated chain of

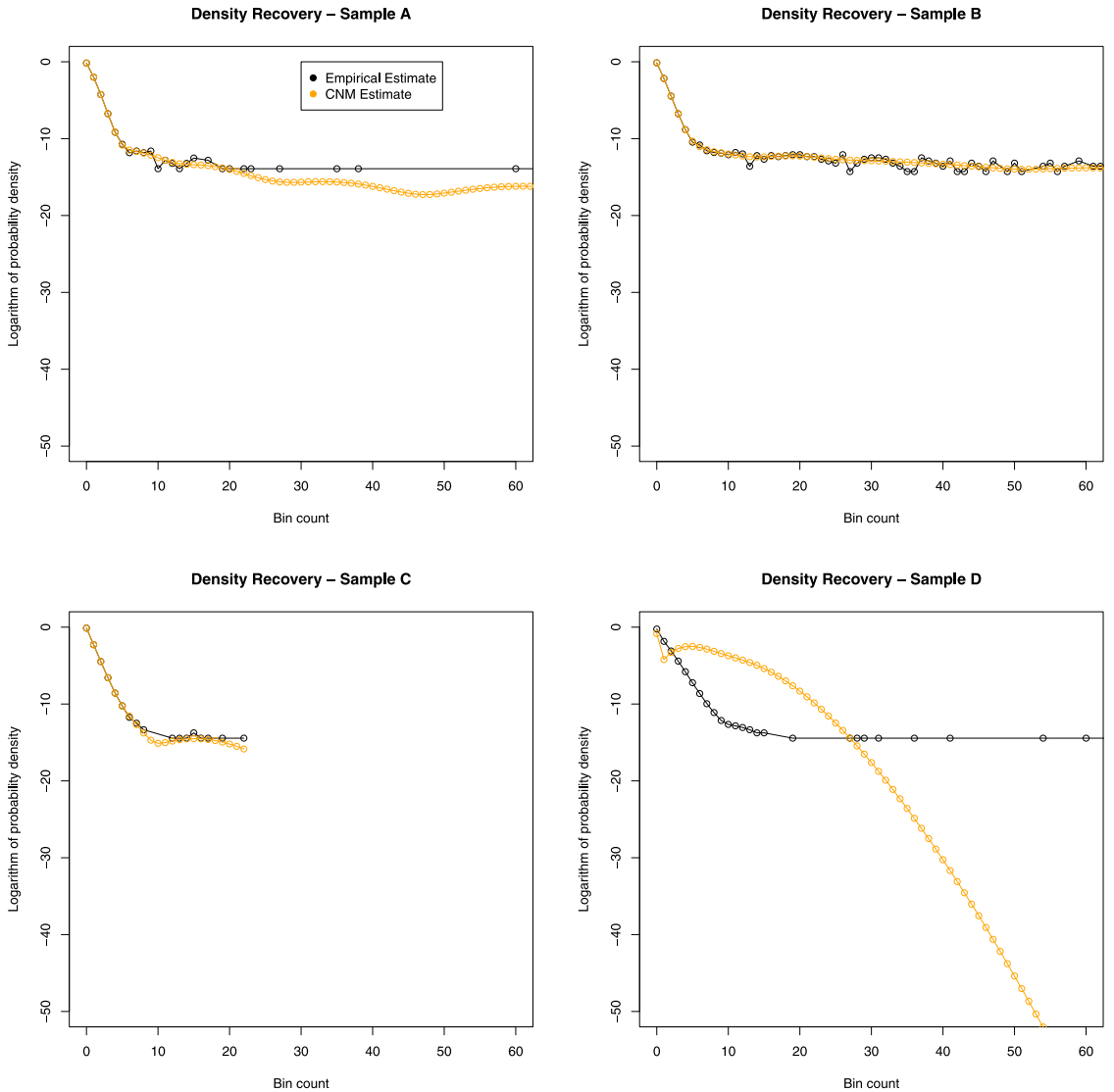

FIGURE 2 | Distribution recovery in each data set. The empirical distribution obtained from the actual data (black line) appears to show a dramatic change in behavior around $x = 9$. While CNM reported failure to converge for samples C and D, it can be seen that sample C's distribution was adequately recovered.

library-preparation procedures, leading to wide variation across hundreds of millions of sites across the genome. Therefore, it is much more logical to model Λ as a continuous random variable.

As such, we adopt the following strategy to assess the fit of various distributions:

1. Fit a continuous mixing distribution $\hat{f}_\Theta(\lambda)$ to the data.
2. Compare $\hat{f}_\Theta(\lambda)$ with the discrete MLE $\hat{f}(\lambda)$, obtained from CNM.

If $\hat{f}_\Theta(\lambda)$ is flexible enough to fit our data, then it should be "similar" to the MLE $\hat{f}(\lambda)$. Thus, we need to compare a discrete distribution, $\hat{f}(\lambda)$, with a continuous distribution, $\hat{f}_\Theta(\lambda)$. Methodology for making such a comparison tends to be *ad hoc*, as the general question is ill-defined.

A common strategy is to smooth out the data through "kernel smoothing"—that is, replacing each point mass in the discrete distribution with an appropriately-scaled kernel distribution, often the normal distribution. The problem with this approach is that the support of the mixing distribution's MLE contains only a handful of points, and the points become dramatically less dense at higher values. Thus, by using this approach, we end up with a distribution that is highly sensitive to the choice of length scale for the kernel.

Instead, we take the following non-parametric approach. Assume that CNM's output [the discrete distribution with support $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_L)$ and associated probabilities $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_L)$] has been generated from a "true" distribution $f(\lambda)$ through the following process:

1. Take some partition $\boldsymbol{P} = (P_1 = 0, P_2, P_3, \ldots, P_L, \infty)$ of the interval $[0, \infty]$.
2. For each $i \in \{1, \ldots, L\}$, replace the probability mass of $f(\lambda)$ that lies in the interval $[P_i, P_{i+1})$ with an equivalent point mass of size $\pi_i = \int_{P_i}^{P_{i+1}} f(u)du$ placed at any point θ_i within that interval.

Now, consider the true distribution's CDF, $F(\lambda) = \int_0^{\lambda} f(u)du$. For $i > 1$, we have:

$$F(P_i) = \int_0^{P_i} f(u)du = \sum_{j=1}^{i-1} \int_{P_j}^{P_{j+1}} f(u)du = \sum_{j=1}^{i-1} \pi_j \quad (1)$$

For the case $i = 1$, we set

$$F(P_1) = F(0) = 0 \quad (2)$$

Note that Equation 2 assumes that Λ is not zero-inflated.

We can now place bounds on the value of $F(\theta_i)$. θ_i must lie in $[P_i, P_{i+1})$, by assumption. Therefore, $F(\theta_i)$ must lie in $[F(P_i), F(P_{i+1}))$, since $F(\lambda)$ is a CDF and is therefore an increasing function of λ. Thus, for $i > 1$, by Equation (1) we have

$$\sum_{j=1}^{i-1} \pi_j \leq F(\theta_i) < \sum_{j=1}^{i} \pi_j \quad (3)$$

and for $i = 1$, we have

$$0 \leq F(\theta_1) < \pi_1 \quad (4)$$

We can use these upper and lower bounds to assess the fit of various candidate mixing distributions to the observed mixing distribution.

Figure 3 shows an example of this procedure, as applied to sample A. We see that all of the candidate mixing distributions considered thus far violate the CNM bounds early on, indicating that these distributions cannot cope with large counts. This motivates the selection of a mixing distribution that can stay within the bounds.

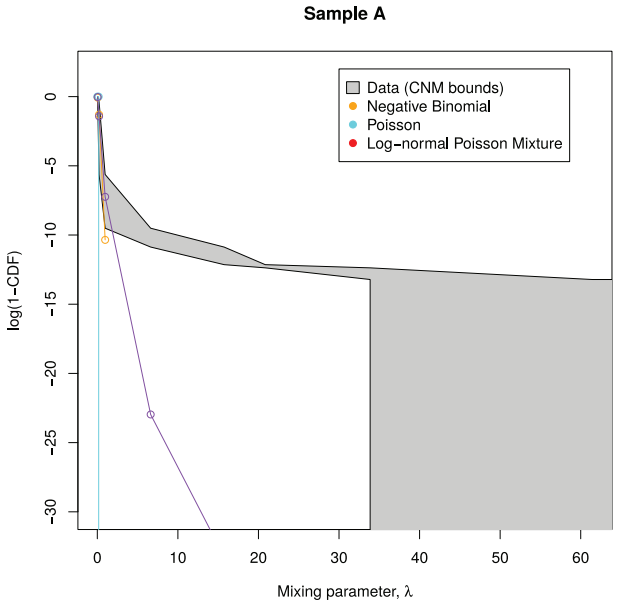

FIGURE 3 | Candidate mixing distributions, and their consistency with the CNM-derived mixing distribution. For clarity, we plot $\log(1 - CDF)$ where *CDF* is the Cumulative Density Function of Λ, and values are calculated only at CNM's λ support points. A mixing distribution that is consistent with CNM's predicted mixing distribution, as derived in Section 3.2, would have a line contained within the shaded region, with the black lines representing upper and lower bounds. Here, the lines do not stay within the bounds, indicating that all of the models deviate from CNM's prediction.

3.3. NB-NB MIXTURE

The curves plotted in **Figure 3** have insufficient curvature to accommodate the sharp turn present in the shaded region. This suggests that a mixture of two distributions is required—consistent with the abrupt change in behavior seen in **Figures 1**, **2**.

We consider the case where X is a mixture of two NB distributions, equivalent to modeling Λ as a mixture of two gamma distributions. In this case, we have a mixing parameter τ, and two separate NB parametrizations (μ_1, r_1) and (μ_2, r_2). Thus,

$$Z \sim Bernoulli(\tau)$$

$$X \sim \begin{cases} NB(\mu_1, r_1) & (Z = 0) \\ NB(\mu_2, r_2) & (Z = 1) \end{cases}$$

We could find the MLE of the parameters with the EM algorithm (Dempster et al., 1977). However, since the EM algorithm can converge very slowly, we accelerated the process using SQUAREM (Varadhan and Roland, 2008). SQUAREM assumes that each step of the EM algorithm can be approximated using a particular quadratic form, allowing us to estimate the cumulation of a large number of EM updates in one go.

Note that we did not consider mixtures of Poisson distributions, since these cases are attended to by the CNM algorithm, and we did not consider mixtures of log-normal Poisson distributions due to the computational complexity.

4. RESULTS

4.1. MODEL FITS

We considered the following distributions:

Count distribution, $f(x)$	Parameters	Mixing distribution, $f(\lambda)$
Poisson	1	Single point
Negative Binomial (NB)	2	Gamma
Log-normal Poisson	2	Log-normal
NB-NB mixture	5	Gamma-gamma mixture

We visually inspect the fits of the various distributions to the full data in **Figure 4**. To quantify the fit, we used Total Variation distance: $d_{TV}(f, g) = \frac{1}{2} \sum_x |f(x) - g(x)|$. The results for this are given in **Figure 5**.

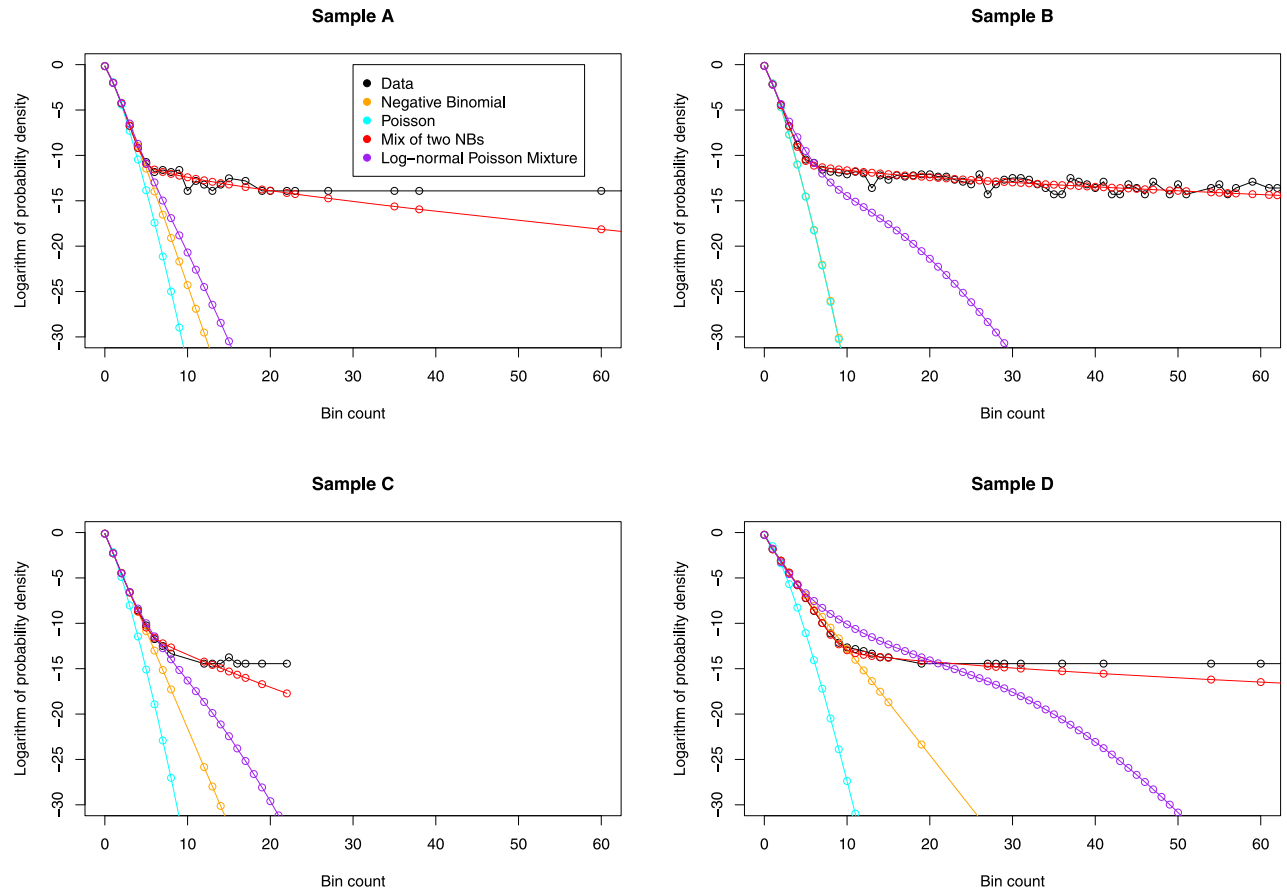

FIGURE 4 | The fit of various distributions to the count data. We again see a striking change of behavior, occurring at around $x = 9$. The only distribution that can model this change is the NB-NB mixture.

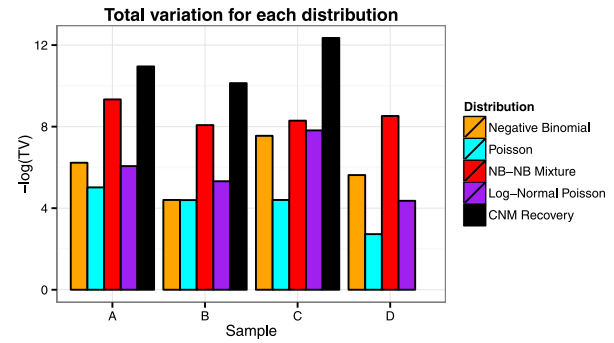

FIGURE 5 | Total variation. We see from the plot that the NB-NB mixture outperforms all of the other distributions, achieving the closest fit to the CNM estimate in the cases where it exists.

Then we examine the similarity between the mixing distributions used and the observed mixing distribution as calculated by CNM, in **Figure 6**.

The NB-NB mixture is the only choice of distribution that can consistently model the higher counts.

4.2. ZERO INFLATION

Zero-inflated models have been proposed to improve model fit in ChIP-seq data (Rashid et al., 2011; Diaz et al., 2012). A zero-inflated model is one such that, with some probability ν, we set $X = 0$. Otherwise, we draw X from a candidate distribution as before.

Having accounted for some of the variation in our data with the NB-NB mixture model, we investigated whether or not zero-inflation can improve model fit further. To quantify this, we delete some proportion of zeros from the data, fit each model, then assess the fit with Total Variation as defined in Section 4.1. We also considered the case where we increase the number of zeros. Note that the log-normal Poisson fit was excluded due to computational complexity.

The results are given in **Figure 7**. For sample A (dog) we see evidence of zero inflation, perhaps due to a less-established genome assembly. In samples B-D, the NB-NB models showed no evidence of zero-inflation. However, when using distributions that cannot account for the heavy tail in the high bin-counts, there is an erroneous indication that a zero-inflation component is necessary.

The general problem of inferring the mixture distribution whilst accounting for zero-inflation is difficult, though similar

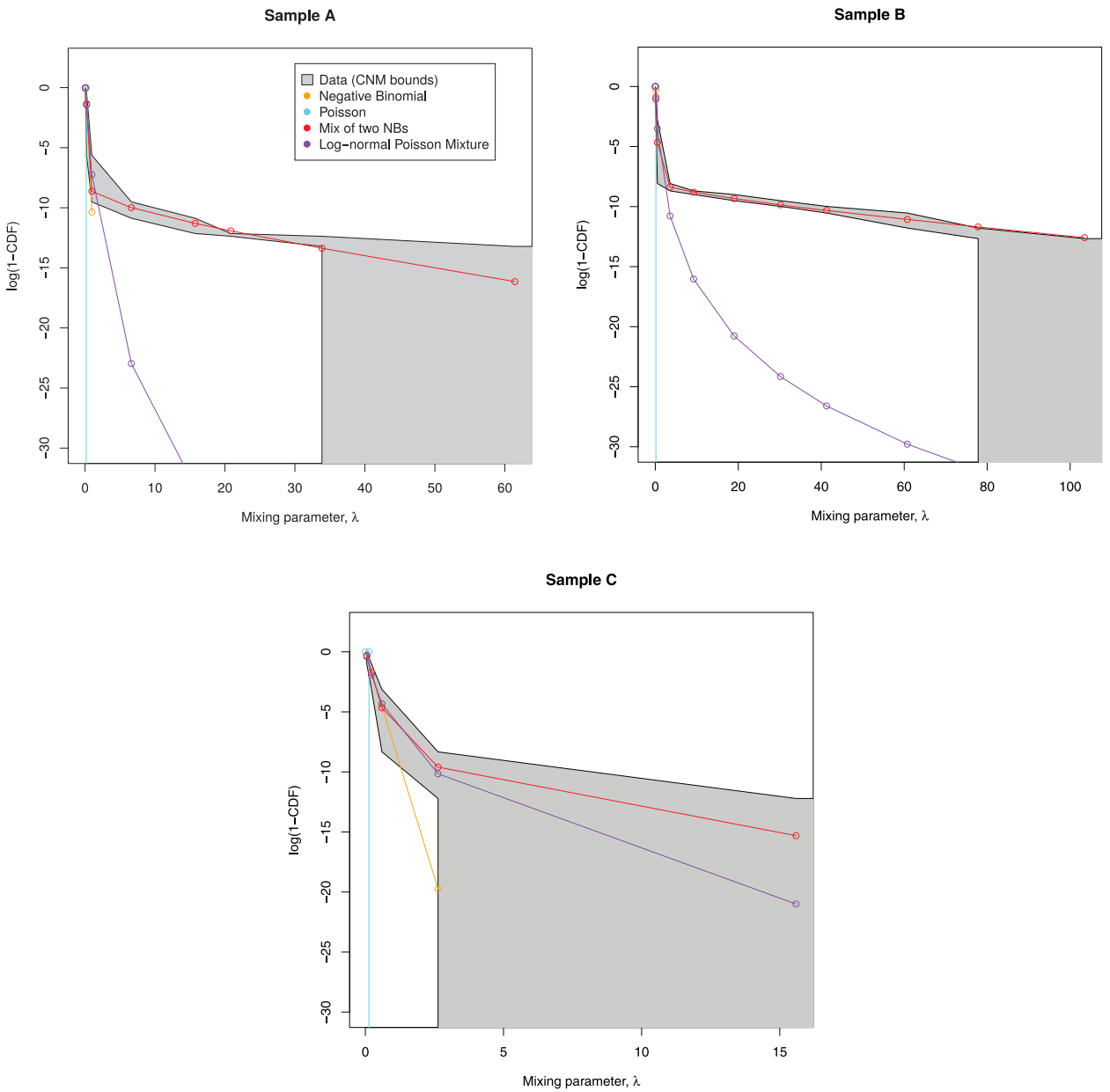

FIGURE 6 | Predicted mixing distributions, as per Figure 3, for each sample with the NB-NB mixture added. (Sample D is omitted since CNM did not recover its distribution.) The NB-NB mixture is the only model whose mixing distribution consistently fits between the CNM bounds.

approaches exist (Lim et al., 2014; Morgan et al., 2014). We envisage an altered version of CNM that accounts for zero-inflation by adding the parameter ν defined above, with its own update step. Additional constraints may be required to make this model identifiable.

5. DISCUSSION

We found that, of the distributions tested, a mixture of two NB distributions best modeled counts in input data, even after removing problematic centromeric and telomeric regions. This result, and the observation that the empirical distribution changes behavior at high counts, reflected two apparent sources of counts in the data—one large population of counts, and another smaller population of higher counts. The regions with higher counts have higher variance than lower counts, and could be mistaken for peaks if using a naïve peak-calling method.

Importantly, we saw that the heavy tail of large counts could cause inadequate models to suggest the need for a zero-inflation component. Our results suggest that researchers should check that large counts are not distorting model fit before assuming that a zero-inflation component is necessary.

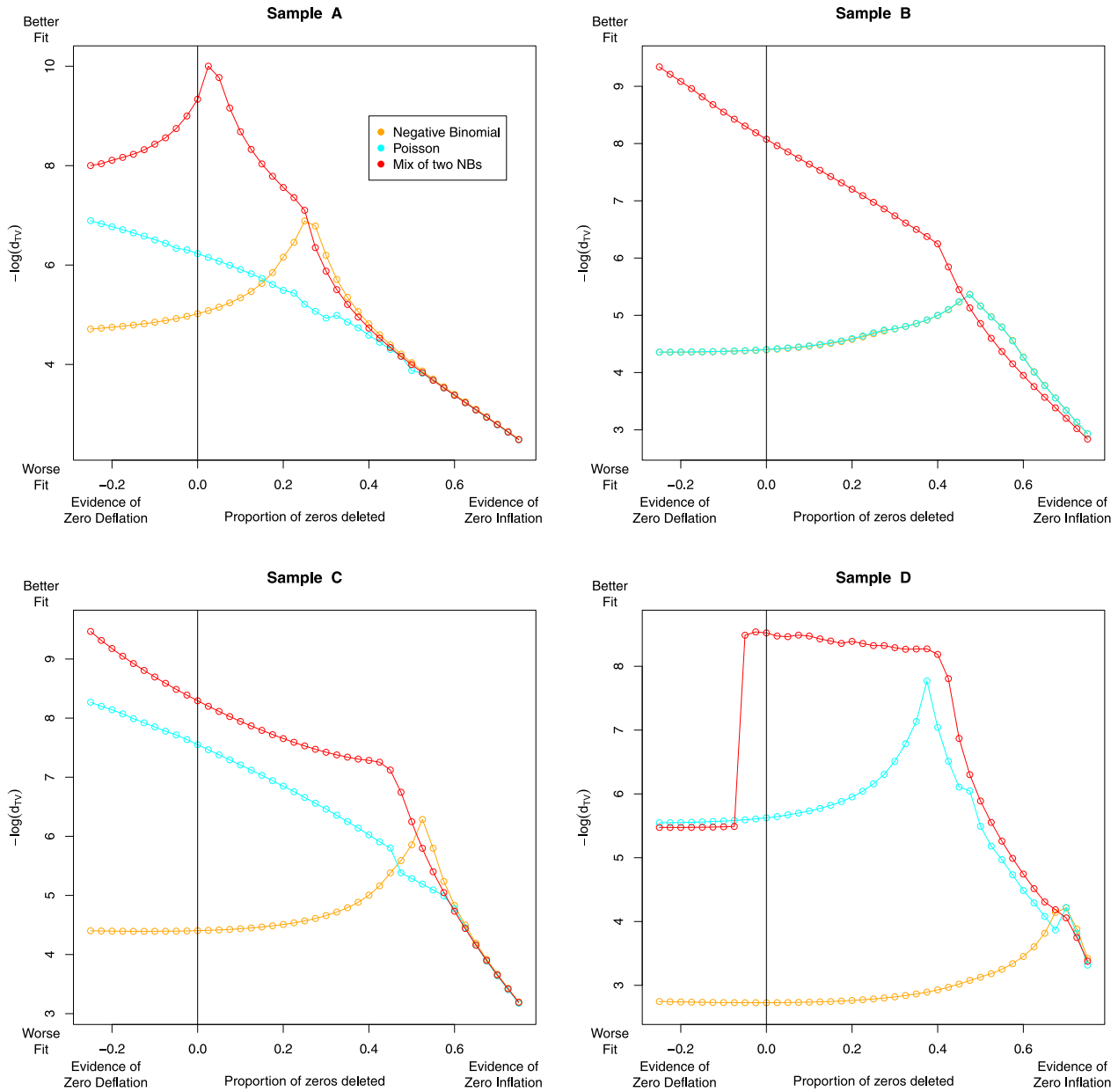

FIGURE 7 | The effect of zero removal on the model fit. Each Subfigure represents an individual sample. A proportion of zeros are removed (or added), and after refitting each model, we reassess model fit with the Total Variation distance d_{TV}, plotted as $-log(d_{TV})$ on the *Y*-axis against zero proportion on the *X*-axis. For severely zero-inflated data, deleting zeros should improve model fit, so we expect to see a "peak" somewhere to the right of the black vertical line. The NB and Poisson models recommend the removal of most of the zero-valued data, a conclusion that seems biologically implausible. In contrast, the NB-NB mixture consistently recommends either no zero-removal, or a modest level of zero-inflation in the case of sample A. Note the dramatic change in the NB-NB mixture's fit in sample D as additional counts are added—this could be due to the accelerated EM algorithm encountering alternative local maxima.

Several aspects of experimental design could influence the abundance of large counts. Any biases in mapping that cause reads to align preferentially to the same locus could cause high-count artifacts. Thus, we might be able to reduce the abundance of large counts by improving mappability (that is, reducing the number of ambiguously-mapped reads). For example, better genome assembly, or use of a better aligner, or using longer reads could all accomplish this task.

Appropriate mixture modeling could apply when simulating ChIP-seq data—a simulation paradigm that fails to account for these different populations of counts cannot test peak-callers properly and overestimates their performance (Zhang et al., 2008b). A better understanding of the properties of the underlying mixture model allows us to simulate noise in ChIP-seq data sets more accurately.

Mixture modeling is also of importance when peak-calling in ChIP-seq data, allowing us to model large counts without removing them. Models that regress on genomic covariates, such as copy number or GC content, can help us explain some of the variation in the data (Rashid et al., 2011; Robinson et al., 2012). However, our unsupervised approach can model the noise, even in the absence of appropriate covariates to regress over—for example, if we have samples that are from less well-elucidated species or that have abnormal copy number events (such as after chromothrypsis). We may also be able to extend the approach to make inferences about the variation that remains after regressing out known covariates. Additionally, we note that Rashid et al. (2011) assume constant dispersion due to general linear modeling restrictions—in contrast, a mixture model permits multiple dispersions.

Alternatively, should we wish to adopt a blacklist strategy, we can construct a blacklist *de novo* by classifying bins into artifact and non-artifact regions (for example, by finding the probability that a bin belongs to each of the NB components in our mixture model). Again, this could be particularly useful for abnormal samples or species.

An alternative approach to smoothing the MLE $\hat{f}(\lambda)$ is to enforce continuity of $\hat{f}(\lambda)$ during maximum likelihood estimation. For example, Liu et al. (2009) minimize a penalized log-likelihood:

$$\ell_p(f) = \ell(f) - \alpha \int_{\lambda_0}^{\lambda_1} \left[f''(\lambda)\right]^2 d\lambda$$

where α, the smoothing parameter, controls how smooth the output function is required to be.

The methods we described have general applicability to other sequencing experiments based on genomic DNA, and not just ChIP-seq. As we increasingly see the emergence of large-scale epigenetic studies based on gDNA assays like ChIP-seq, it is important to choose models whose false discovery rates are robust. Properly accounting for the null variability in ChIP-seq data is vital to avoid false positives.

FUNDING

We acknowledge the support of the University of Cambridge, the Medical Research Council, Cancer Research UK, and Hutchison-Whampoa.

ACKNOWLEDGMENTS

We are grateful to Caryn Ross-Innes, Jason Carroll, Mike Wilson and Duncan Odom for access to data, to Mike L. Smith for data hosting, and to John Marioni and Mark Robinson for helpful discussions.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: http://www.frontiersin.org/journal/10.3389/fgene.2014.00399/abstract

REFERENCES

Boes, D. C. (1966). On the estimation of mixing distributions. *Ann. Math. Stat.* 37, 177–188. doi: 10.1214/aoms/1177699607

Cairns, J., Lynch, A. G., and Tavaré, S. (2013). "Statistical aspects of ChIP-seq analysis," in *Advances in Statistical Bioinformatics*, Chapter 7, eds K.-A. Do, Z. S. Qin, and M. Vannucci (New York, NY: Cambridge University Press), 138–169. doi: 10.1017/CBO9781139226448.008

Cairns, J., Spyrou, C., Stark, R., Smith, M. L., Lynch, A. G., and Tavaré, S. (2011). BayesPeak - an R package for analysing ChIP-seq data. *Bioinformatics* 27, 713–714. doi: 10.1093/bioinformatics/btq685

Carroll, T. S., Liang, Z., Salama, R., Stark, R., and de Santiago, I. (2014). Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. *Front. Genet.* 5:75. doi: 10.3389/fgene.2014.00075

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B* 39, 1–38.

Diaz, A., Park, K., Lim, D. A., and Song, J. S. (2012). Normalization, bias correction, and peak calling for ChIP-seq. *Stat. Appl. Genet. Molec. Biol.* 11:9. doi: 10.1515/1544-6115.1750

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. *J. Am. Stat. Assoc.* 73, 805. doi: 10.1080/01621459.1978.10480103

Landt, S. G., Marinov, G. K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., et al. (2012). ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* 22, 1813–1831. doi: 10.1101/gr.136184.111

Lim, H. K., Li, W. K., and Yu, P. L. (2014). Zero-inflated Poisson regression mixture model. *Comput. Stat. Data Anal.* 71, 151–158. doi: 10.1016/j.csda.2013.06.021

Liu, L., Levine, M., and Zhu, Y. (2009). A functional EM algorithm for mixing density estimation via nonparametric penalized likelihood maximization. *J. Comp. Graph. Stat.* 18, 481–504. doi: 10.1198/jcgs.2009.07111

Lunn, D. J., Thomas, A., Best, N., and Spiegelhalter, D. (2000). WinBUGS - A Bayesian modelling framework: concepts, structure, and extensibility. *Stat. Comput.* 10, 325–337. doi: 10.1023/A:1008929526011

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008). RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.* 18, 1509–1517. doi: 10.1101/gr.079558.108

Morgan, C. J., Lenzenweger, M. F., Rubin, D. B., and Levy, D. L. (2014). A hierarchical finite mixture model that accommodates zero-inflated counts, non-independence, and heterogeneity. *Stat. Med.* 33, 2238–2250. doi: 10.1002/sim.6091

Myers, R. M., Stamatoyannopoulos, J., Snyder, M., Dunham, I., Hardison, R. C., Bernstein, B. E., et al. (2011). A user's guide to the Encyclopedia Of DNA Elements (ENCODE). *PLoS Biol.* 9:e1001046. doi: 10.1371/journal.pbio.1001046

Park, P. J. (2009). ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* 10, 669–680. doi: 10.1038/nrg2641

Rashid, N. U., Giresi, P. G., Ibrahim, J. G., Sun, W., and Lieb, J. D. (2011). ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. *Genome Biol.* 12:R67. doi: 10.1186/gb-2011-12-7-r67

Robinson, M. D., Strbenac, D., Stirzaker, C., Statham, A. L., Song, J., Speed, T. P., et al. (2012). Copy-number-aware differential analysis of quantitative DNA sequencing data. *Genome Res.* 22, 2489–2496. doi: 10.1101/gr.139055.112

Ross-Innes, C. S., Stark, R., Teschendorff, A. E., Holmes, K. A., Ali, H. R., Dunning, M. J., et al. (2012). Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature* 481, 389–393. doi: 10.1038/nature10730

Roueff, F., and Rydén, T. (2005). Nonparametric estimation of mixing densities for discrete distributions. *Ann. Stat.* 33, 2066–2108. doi: 10.1214/009053605000000381

Saha, K., and Paul, S. (2005). Bias-corrected maximum likelihood estimator of the negative binomial dispersion parameter. *Biometrics* 61, 179–185. doi: 10.1111/j.0006-341X.2005.030833.x

Schmidt, D., Wilson, M. D., Ballester, B., Schwalie, P. C., Brown, G. D., Marshall, A., et al. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. *Science* 328, 1036–1040. doi: 10.1126/science.1186176

Simar, L. (1976). Maximum likelihood estimation of a compound Poisson process. *Ann. Stat.* 4, 1200–1209. doi: 10.1214/aos/1176343651

Song, Q., and Smith, A. D. (2011). Identifying dispersed epigenomic domains from ChIP-Seq data. *Bioinformatics* 27, 870–871. doi: 10.1093/bioinformatics/btr030

Spyrou, C., Stark, R., Lynch, A. G., and Tavaré, S. (2009). BayesPeak: bayesian analysis of ChIP-seq data. *BMC Bioinform.* 10:299. doi: 10.1186/1471-2105-10-299

Thygesen, H. H., and Zwinderman, A. H. (2006). Modeling SAGE data with a truncated gamma-Poisson model. *BMC Bioinform.* 7:157. doi: 10.1186/1471-2105-7-157

Tucker, H. (1963). An estimate of the compounding distribution of a compound Poisson distribution. *Theor. Probab. Appl.* 8:195. doi: 10.1137/1108021

Varadhan, R., and Roland, C. (2008). Simple and globally convergent methods for accelerating the convergence of any EM algorithm. *Scand. J. Stat.* 35, 335–353. doi: 10.1111/j.1467-9469.2007.00585.x

Venables, W. N., and Ripley, B. D. (2002). *Modern Applied Statistics with S, 4th Edn*. New York, NY: Springer. doi: 10.1007/978-0-387-21706-2

Wang, Y. (2007). On fast computation of the non-parametric maximum likelihood estimate of a mixing distribution. *J. Roy. Stat. Soc. B* 69, 185–198. doi: 10.1111/j.1467-9868.2007.00583.x

Wu, S., Wang, J., Zhao, W., Pounds, S., and Cheng, C. (2010). ChIP-PaM: an algorithm to identify protein-DNA interaction using ChIP-Seq data. *Theor. Biol. Med. Model.* 7:18. doi: 10.1186/1742-4682-7-18

Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., et al. (2008a). Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* 9, R137. doi: 10.1186/gb-2008-9-9-r137

Zhang, Z. D., Rozowsky, J., Snyder, M., Chang, J., and Gerstein, M. (2008b). Modeling ChIP sequencing *in silico* with applications. *PLoS Comp. Biol.* 4:e1000158. doi: 10.1371/journal.pcbi.1000158

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Received: 09 June 2014; accepted: 29 October 2014; published online: 14 November 2014.

*Citation: Cairns J, Lynch AG and Tavaré S (2014) Quantifying the impact of inter-site heterogeneity on the distribution of ChIP-seq data. Front. Genet. **5**:399. doi: 10.3389/fgene.2014.00399*

This article was submitted to Bioinformatics and Computational Biology, a section of the journal Frontiers in Genetics.